Published in final edited form as:

Socioecon Plann Sci. 2017 March ; 57: 1–13. doi:10.1016/j.seps.2016.09.005.

Cross-Sectional Surveys: Inferring Total Eventual Time in Current State Using Only Elapsed Time-to-Date

Richard C. Larson

Abstract

We focus on snapshot surveying of sub-populations whose members are in a temporary state and where one of the questions asked is the elapsed time already spent in that state. From these answers we develop probabilistic and statistical procedures to estimate the distribution of **total time** that will eventually be spent in that state by any random individual who enters the state. The method relies on a selection bias often found in temporal sampling, sometimes called “random incidence” or “longevity bias.” We develop results for several types of sampling, including random and fixed times of surveying, random and fixed times of entering the state, and sampling only those who have already spent some minimal specified time in the targeted state. An example with post-doc data is included to demonstrate the steps.

Keywords

Cross-sectional survey; snapshot survey; career duration; random sampling; post-doc careers

Often in survey research we wish to poll at one point in time a random selection of individuals who are in some temporary state. Such polls are often called “cross-sectional surveys” or “snapshot surveys.” While the inspiration of this paper is the surveying of post-doctoral fellows (“post-docs”), such random surveys are seen in myriad domains, especially in services sectors, both public and private. The public sector has many needs for this type of research, investigating for instance durations of criminal careers, queue waits for medical treatment of U.S. veterans, waits for placement in public housing, career lengths of public school teachers, just to name a few. In the private sector, we may be surveying automobile owners and homeowners, - for durations of ownerships, individuals with mortgages, patients receiving chemotherapy, etc. Representative literature is found in marketing (Rindfleisch et. al 2008), health care (Barnett et. al. 2012), education (Kaia, 2006; Tuncalp, et. al. 2014), banking (Voleti & Bharadwaj 2015), transportation (Oum 1979), and criminal justice (Blumstein & Cohen 1987).

Our focus is those cross-sectional (“one-shot”) surveys that have duration as a key concern, i.e., the total amount of time that the surveyed individuals will be in the temporary state that defines the sub-population. In our domain-specific development below, this time is the

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

duration of post-doc status, often varying in practice from less than one year up to ten years or more! In the survey we ask how long the individual surveyed has been in the targeted state. The goal of the paper is to use those responses, together with the ‘physics’ of the sampling process, to infer the probability distribution of the entire amount of time that all members of the sub-population will be in that state, summing past and future times in the state.

The mathematical setting is framed with the arguments for “random incidence” found in *Urban Operations Research* (Larson, Odoni, 1981, Sec. 2.13). Many of these ideas are also presented in Drake (1967) and build from some basic properties of renewal processes (Cox, 1962). But the finite-length durations that we are examining are not associated with renewal processes, as renewal processes involve independent recurrent events that occur repeatedly over time. In our modeled environment, members of the sub-population enter the targeted state once and remain there for a finite length of time. We only assume statistical independence between and among members of the surveyed sub-population. We also refer the reader to a foundational paper by Kaplan (1997) where he leverages the type of longevity bias we exploit within “random incidence” to examine cases in which two types of events occur to members of a sub-population: a trigger event and then perhaps a related follow-on event. He applies his results to a number of domains, including HIV/AIDS, employment, poverty and criminal careers.

What we call “longevity bias” has been considered and discussed in many domains, usually with the goal of “correcting the bias.” One example, seen often in the medical literature, is called “*length-biased sampling*,” the ‘problem’ being that “...the sample mean becomes a biased estimator of the population mean.” (Academy Health, 2015) “*Current status bias*” is another phrase sometimes used to describe a net policy-inferential property stemming from longevity bias. For example, Phelan and Link (1998) argue that studies characterized by current status bias depict “poor people, homeless people welfare recipients, mental hospital patients and criminal offenders as hard-core groups, chronically mired in an undesirable condition largely because of deviant attributes that appear to characterize the group.” They argue that the bias underestimates the numbers who enter such undesirable states for short periods of time and then successfully leave. As a result, the chronic inhabitants of these states, such individuals over-sampled by current status bias, inappropriately guide public perceptions and ameliorative policies.

Our focus is not *current status bias*, nor is it on “correcting” longevity bias. Rather, we use this natural bias to derive useful new results relating to the total time that a random individual spends in the temporary state, given only that he/she enters that state, unaffected by any longevity bias.

A companion paper (Andalib, Ghaffarzadegan, Larson; 2016) applies these methods to a mathematical depiction of the post-doc population as a queueing system with reneging, the early departures due to system saturation as there are far fewer new tenure-track positions available each year than there are new incoming post-docs seeking such positions. A second companion paper is mentioned at the end of this paper.

For the remainder of this paper, we use post-docs as the application environment for the results, which are general. In that context, an alternative title for the paper could be, *“Finding the Total Time Spent as Post-docs Using a Random Sample of Current Post-docs.”* We choose to refer to post-docs rather than some vague “temporary state of an individual” for two reasons: (1) to reflect the founding motivation for this paper (supported by a workforce-focused research grant from the NIH), and (2) to assist the reader visualize the problem and approach. The slowly growing number of post-docs nationally has increasingly been a focus of national research policy. The National Science Foundation (NSF), the National Institutes of Health (NIH) and others periodically poll post-docs to obtain useful information about their activities and durations of their post-doc careers. The plight of post-docs is very much a public sector problem, as the bulk of their research funding comes from public funds and a large fraction of post-docs hold their appointments in public universities. Our primary focus is the duration of the post-docs careers. We analyze this via methods of applied probability.

We are interested in the probability density function (PDF) of the random variable W , the total duration of a post-doc career for a randomly starting post-doc. Surveys of active post-docs cannot ask for this total time as post-doc, as very few post-docs know precisely when they will leave the post-doc status. Instead, such surveys ask for the length of time that a randomly chosen post-doc has been a post-doc, i.e., the elapsed time to date as a post-doc. The key question is this: Can we infer from these reported elapsed times the distribution of *total time* that post-docs spend in the post-doc status? For reader convenience a summary of technical terminology is provided in Table 1.

1. Visualizing the Sampling Situation

Consider a picture of our sampling physics, depicted in Figure 1. The time axis, moving from past to future, left to right, is shown at the bottom of the figure. Each of the blue horizontal lines represents one post-doc career, from start time to end time, as illustratively labeled for the top-most career in the figure. Note that post-doc careers start at different times, end at different times and have varied durations.

We now assume that a random survey of post-docs occurs at the time shown in the figure. The vertical dotted line extending upward from this sampling time is shown intercepting some of the post-doc careers and missing others. A post-doc whose career is intercepted in this way is an active post-doc and will be included in our sample. Post-docs who are missed by the vertical line are excluded from our sample since either their career has already ended or it has not yet started, examples shown in the figure. In general, the longer careers have a greater chance of being intercepted than shorter ones. We must consider and take analytical advantage of this sampling bias towards careers of longer duration. Our key data measurement is reported time-to-date as an active post-doc. Our ultimate goal is to use the information we obtain to estimate the probability distribution for the duration of the entire post-doc career, for all post-docs, using only the “snap-shot” sampled data reporting time-to-date as a post-doc. As we develop the mathematics for each of our three sampling situations, we employ a “Temporal Sampling Diagram” depicting a fully labeled characterization of the situation, building from the temporal relationships shown in Figure 1.

In all of our work we assume a type of “steady state” working of the system, with the total number of post-docs generated each year neither growing nor declining appreciatively and durations of post-doc careers not changing significantly over time.

The perceptive reader may note that the steady state assumption is at odds with the recent secular growth nationally in the number of post-docs. It is true that steady state assumes no trending growth or decline in numbers of post-docs. However, over the typical career length of a post-doc, usually two to five years, the national growth may not even be perceptible as a trend, due to random fluctuations in annual figures, so the steady state assumption can be applied as a reasonable approximation.

We first discuss the timings of the survey and the start of the post-doc periods. The easiest case is when the post-doc careers commence at random times during the year, with equal starting rates throughout the year, and the survey is conducted at a given time during the year, independent of the status of any post-doc(s). The mathematical model we develop below in Sec. 2 is compatible with these assumptions. For example, post-doc careers could commence at random times while the survey is conducted at a pre-scheduled time in, say, October.

How might one model the random start times of post-doc careers? One possibility is that a homogeneous Poisson process could generate the start times with rate λ new post-doc careers started per unit time. As an example, suppose that the Poisson rate parameter λ is 100 new post-docs per year and that all post-doc careers are precisely three years in duration. Then, on average in steady state, at any given time there would be a random number of active post-docs in the system, the random variable having a Poisson distribution with mean 300 active post-docs, about 100 from each of the previous three years. But a Poisson career generation process is not required nor is a fixed duration post-doc career. We will not model explicitly the post-doc career generation process and we will allow the duration of the post-doc career to have any arbitrary probability distribution.

In Sec. 3 we reconsider the problem of Sec. 2, but for a situation in which sampled post-docs must have been in that status for a given (minimal) period of time. In popular parlance, “newbies” are excluded!

In Sec. 4 we consider the case when **both** post-doc initiation date and survey date are fixed, and this requires significant modifications to our earlier results.

2. Case 1: Random Career Start Times, Fixed Survey Time

Recall that the random variable W is defined to be the total duration of a post-doc career for a randomly starting post-doc. Now consider the random variable X , defined to be the elapsed post-doc time to date of a random sampled individual. The relationships are shown in the diagram of Figure 2¹.

¹Everything we develop about the random variable X , defined to be the elapsed time from start of the post-doc period until the time of sampling, also applies to the random variable $Y = W - X$, defined to be the remaining time in the post-doc interval from the time of sampling until termination of post-doc status. X and Y have identical probability laws.

We seek to extract an estimate of the PDF of career length W from the sampled survey data on elapsed career time to date, X . We assume randomness of post-doc start dates, with the survey date either fixed or random.

We now proceed with the analysis. Via the survey we are entering post-doc careers at random times. Everything else being equal, longer careers are more likely to be sampled than shorter careers. The key observation is this: The probability that any particular post-doc will be interviewed is directly proportional to two quantities: (1) the relative frequency of post-docs whose post-doc total career duration W is in a specified incremental window (w to $w + dw$) and (2) the post-doc's total career time w as a post-doc. (We use the notation that any random variable is given an upper case letter, e.g. W , and its numerical experimental value appears as the corresponding lower case letter, w .) The favoring of longer careers is the selection bias sometimes called "random incidence." With regard to (1), if there are twice as many 2-year duration post-docs as 4-year duration post-docs, there is a factor-of-2 bias towards sampling two-year duration post-docs. With regard to (2), a post-doc who spends four years as a post-doc would be twice as likely to be sampled as one who spends only two years. But both proportionality factors work together in unison. For example, if there are twice as many post-docs whose post-doc career duration is two years in contrast to those for whom it is four years, then the relative frequency factor favors the two-year post-docs twice as much as the four-year post-docs. But the fact that the four-year post-docs spend twice as much time being a post-doc as the two-year post-docs places a factor-of-2 favoritism towards the four-year post-docs. The net result: For this illustrative example, our next sampled post-doc is equally likely to be a two-year post-doc or a four-year post-doc.

We now formalize these observations. First we consider the probability Pdw of selecting a post-doc whose post-doc career duration is in the interval $[w, w + dw]$. We know that this probability is proportional to the relative frequency of such durations, namely $f_W(w)dw$, and to the length of the career w . So, we can write

$$Pdw = \alpha w f_W(w) dw, \quad (1)$$

where α is a normalization constant. Since Eq. (1) must be a proper PDF, it must integrate to 1 over $[0, \infty)$, and thus the normalizing factor is $E[W]$, the mean duration of a post-doc career,

$$Pdw = \frac{w f_W(w) dw}{E[W]} \quad w \geq 0 \quad (2)$$

Now suppose we sample an individual post-doc whose eventual total time as a post-doc is w . We are equally likely to perform our sampling in any small t interval over her entire post-doc career duration. Hence we know that the elapsed time x already spent as a post-doc is a random variable uniformly distributed over the interval $[0, w]$. More formally, if $f_{X|W}(x|w)dx$ is the conditional probability that $x \in X \in x + dx$, given that $W = w$, then

$$f_{X|W}(x|w) = \frac{1}{w}, \quad 0 \leq x \leq w. \quad (3)$$

We can now write an expression for $f_X(x)dx$, the unconditional probability that the elapsed time to date as a post-doc of a randomly selected post-doc assumes a value between x and $x+dx$,

$$f_X(x)dx = \left[\int_{w=x}^{\infty} \frac{1}{w} \frac{w f_W(w) dw}{E[W]} \right] dx = \left[\int_{w=x}^{\infty} \frac{f_W(w) dw}{E[W]} \right] dx = \frac{1 - F_W(x)}{E[W]} dx, \quad (4)$$

where $F_W(x)$ is defined to be the cumulative distribution function (CDF) for the random variable W , namely,

$$F_W(x) \equiv \Pr\{W \leq x\}.$$

For any non-negative random variable such as W , the CDF is a monotonically non-decreasing function for all values of its argument. The CDF starts at $F_W(0) = 0$, climbs to higher values for increasing values of its argument, and eventually reaches 1.0: $F_W(\infty) = 1$. The numerator term $[1 - F_W(x)]$ on the right hand side of Eq.(4) is the complementary cumulative distribution function (CCDF), defined to be the probability that the corresponding random variable is greater than the argument of the CCDF,

$$[1 - F_W(x)] \equiv \Pr\{W > x\}.$$

The CCDF is a monotonically non-increasing function of its argument, starting at 1.0 and eventually – for large values of its argument – ending at 0. In examining Eq.(4), we see that $f_X(x)$ is directly proportional to the CCDF $[1 - F_W(x)]$. This means that $f_X(x)$ is a monotonically non-increasing function of its argument. In simple terms, this PDF starts at its maximum value at $x = 0$, and eventually declines from there to 0 as x increases.

Example—Suppose we knew that post-doc total durations were negatively exponentially distributed with mean $E[W] = 4$ years. Then, substituting into Eq.(4), we obtain

$$f_X(x)dx = \frac{1 - F_W(x)}{E[W]} dx = \frac{1 - (1 - e^{-x/4})}{4} dx = \frac{1}{4} e^{-x/4} dx, \quad \text{for } x \geq 0.$$

This result, surprising to many, says that if post-doc total career durations are exponentially distributed with mean 4 years, then the elapsed post-doc durations to date of randomly selected post-docs are also exponentially distributed with mean 4 years. For some, this is a

most counter-intuitive result, as we may believe there must be a missing factor of 2 to adjust for the fact that – on average -- we only sampled individuals ‘half way through’ their post-doc careers. The resolution of the apparent dilemma resides in the temporal sampling bias due to random incidence². The factor of 2 is a property of the exponential PDF; other PDF’s could have factors greater or less than 2. Note that $f_X(x)$ is a monotonically decreasing function of x , as we expect.

In examining Eq.(4), the usual interpretation is that $f_W(w)$ is known and $f_X(x)$ is unknown. One typically solves for $f_X(x)$. But our situation is different. We have data that depicts a large sample from $f_X(x)$, and we wish to infer $f_W(w)$. So, in essence, we ‘know’ $f_X(x)$ and need to know $f_W(w)$, or equivalently $F_W(x)$. We can do this by manipulating Eq.(4). First from Eq.(4) we write

$$f_X(x) = \frac{1 - F_W(x)}{E[W]} \quad x \geq 0.$$

Since $E[W]$ is an unknown, we must not use it. But we know that the CCDF for W , namely $[1 - F_W(x)]$, is directly proportional to $f_X(x)$, i.e.,

$$[1 - F_W(x)] \propto f_X(x).$$

This suggests a simple procedure for finding the probability law for the total time W spent as a post-doc. Simply normalize the large sample histogram found for the elapsed time X spent to date in those sampled from the post-doc population. The normalization will make the area under the histogram equal to 1.0, a requirement for any PDF. Then observe that the CCDF for the desired random variable W , $[1 - F_W(w)]$, is directly proportional to normalized histogram. This gives us the shape of the desired CCDF. Since $F_W(x) \equiv \Pr\{W \leq x\}$ and $F_W(0) = \Pr\{W \leq 0\} = 0$, we must have $[1 - F_W(0)] = 1$. This suggests that the proportionality relationship can be rewritten as an equation,

$[1 - F_W(x)] = \frac{f_X(x)}{f_X(0)}$, assuming that $f_X(0) > 0$, which it must be by the monotonicity property of $f_X(x)$. Finally we can write for the CDF,

$$F_W(x) = 1 - \frac{f_X(x)}{f_X(0)} \quad x \geq 0, \quad (5)$$

²A famous alternative form of the paradox is found with Poisson arrival processes. A passenger arrives at a random time at the bus stop, where buses arrive totally unscheduled according to a Poisson process with mean headways between buses of 10 minutes. The passenger asks a traffic policeman there how long it has been since the last bus visited the bus stop, and she answers, “Ten minutes.” Then you wait for the next bus to arrive, and the mean duration is ten minutes. So, you have found that the mean headway between two successive buses is 20 minutes, not ten! The resolution to the apparent paradox is found in random incidence: The arriving passenger – by his random arrival time – tends to favor inter-arrival gaps that are longer than average.

or for the PDF,

$$f_W(x) = -\frac{1}{f_X(0)} \frac{d}{dx}(f_X(x)) \quad x \geq 0. \quad (6)$$

These are the transformations we are looking for: We have derived the probability law for the total time W in a random post-doc career only from the histogram of elapsed times to date of cross-sectional sampled post-docs. As a check for Eq.(5), since $f_X(x)$ has maximum value at $x = 0$ and declines from there, we are assured that $F_W(x)$ is a monotonically non-decreasing function and always between 0 and 1, as required. And as a check for Eq.(6), we

know that $\frac{d}{dx}(f_X(x)) \leq 0$ since $f_X(x)$ is a monotonically non-increasing function of x , so with the minus sign in Eq.(6) we are always assured that $f_W(x) \geq 0$, as required.

The mean total time $E[W]$ spent a post-doc career can be found in any one of three ways. From either Eq.(5) or Eq.(6), we have

$$E[W] = \int_0^\infty [1 - F_W(x)] dx, \quad (7)$$

where we use the fact that the mean of a non-negative random variable is equal to the integral over $[0, \infty)$ of its CCDF. Using Eq.(6), we can write

$$E[W] = \int_0^\infty x F_W(x) dx. \quad (8)$$

From Eq.(4) note that $f_X(0) dx = \frac{1 - F_W(0)}{E[W]} dx = \frac{1}{E[W]} dx$, implying that $f_X(0) = \frac{1}{E[W]}$, giving us a ‘by inspection’ way to find $E[W]$ from the normalized empirical histogram, namely,

$$E[W] = \frac{1}{f_X(0)}. \quad (9)$$

We summarize the action consequences of the derivation in the following algorithm:

1. Considering sampled post-docs from a given population, for the corresponding histogram of time spent to date in as a post-doc, **normalize the histogram** so that it has unit area. This gives it the required unit-area property of a PDF.
2. If the normalized histogram found in (1) contains ‘bumpy’ downward steps, perhaps due to small sample size, one may want to smooth it while keeping its

total area equal to 1.0. We will call the final smoothed histogram $\hat{f}_X(x)$, the ‘hat’ over the f representing a data-estimated PDF.

3. Identify the numerical value for $\hat{f}_X(0)$, which is the maximum value for the function $\hat{f}_X(x)$ for all $x \geq 0$.

4. Find the desired estimated CDF of the random variable W , $\hat{F}_W(x)$, using Eq.(5),

$$\hat{F}_W(x) = 1 - \frac{\hat{f}_X(x)}{\hat{f}_X(0)}.$$

5. Find the desired estimated PDF of the random variable W , $\hat{f}_W(x)$, using Eq.(6),

$$\hat{f}_W(x) = -\frac{1}{\hat{f}_X(0)} \frac{d}{dx}(\hat{f}_X(x)).$$

Due to possible ‘bumpiness’ of the empirical histogram, it is our belief that the CDF will provide the more accurate estimated probability law for W .

6. Possible complication: discretized data. Often, perhaps most of the time, the data will be obtainable only in discrete buckets, not continuously as the theory of PDF’s requires. So we may need to make some adjustments for data found in discrete buckets, especially if there are round-off errors. Our suggestions to deal with these issues are illustrated in the numerical example given at the end of the paper.

3. The Case of Minimum Required Time as Post-doc

Sometimes national surveys may be designed only to include individuals who have been in their temporary state at least for some pre-specified minimal amount of time τ . For example, surveys of new car owners may exclude people who have purchased their car within the past three months, as the survey seeks individuals who have had sufficient time to establish a firm set of beliefs about their cars. The NSF excludes in their classic doctoral surveys PhDs who graduated within the last 18 months³. In the case of new car owners, $\tau = 3$ months; for PhD’s who immediately upon graduation become post-docs, $\tau = 18$ months.

Now let’s consider such a “truncated survey” of post-docs in which we exclude from the sample those who have been post-docs for less than τ time units. In practice this might mean that the survey instrument has as a first question, “How long have you been a post-doc?” If the response is less than τ , then the survey stops and the respondent is thanked. That individual is not included in the final tabulations. All others are included.

We recall that the random variable W corresponds to the eventual total time as a post-doc, *for all post-docs*. Here we will not seek to find the probability law of W , as we will exclude all those who have been post-docs for less than time τ , some of whom by time τ may already have terminated their post-doc status! Rather, we will seek the probability law for a new random variable,

³Some NSF doctoral surveys ask for numerical values of our key variable - elapsed time as post-doc, but only for those whose PhD’s were awarded at least 18 months prior to the survey. This process results in an underestimate of the number of post-docs, as it misses individuals whose PhD’s were awarded less than 18 months prior to the survey and who also became post-doc’s within 18 months of the PhD. We will address these NSF surveys in a companion paper.

$W_\tau \equiv$ total post- doc career duration for a randomly sampled post- doc, given that this individual has been a post- doc for at least time τ , $\tau > 0$.

The corresponding temporal sampling diagram is shown in Figure 3. In the figure we have introduced a new random variable,

$\{W|\tau\} \equiv$ the truncated post- τ total time as a post- doc, where the truncation excludes the first τ months as post- doc, given that the individual has been a post- doc for at least τ months.

From the figure we see the additive relationship tying together the three time quantities,

$$W_\tau \equiv \{W|\tau\} + \tau.$$

Importantly, the random variable W_τ is not the same as W , as W_τ depicts only individuals who have been post-docs for at least τ units of time. The PDF for W will contain probability mass in the interval $[0, \tau)$ whereas the PDF for W_τ will not. Finally, the random variable X is the time-after- τ of sampling of a random post-doc who is retained in the sample. A value of 0 for X corresponds to sampling of a retained post-doc just when that post-doc becomes eligible for this survey, i.e., when she has worked as a post-doc for τ months.

Our task now is to find the probability distribution for $\{W|\tau\}$, that is, the τ -reduced post-doc career duration for those randomly sampled post-docs who have been post-docs for at least τ months. We assume, as in Sec. 2, that we have a large sample size histogram of the elapsed time of post-docs for sampled post-docs who have worked in that capacity for at least τ months. We need to be careful with the data, as the post-docs responding to the survey will report back their full time as post-docs to date. As we will show, we will be working with that reported number, but initially with τ time units subtracted out. We will add back the τ time units at the end of the analysis.

We now show that we can reduce this problem to the problem previously solved in Sec. 2. By arguments of Sec. 2 of this paper, we see that X is uniformly distributed over the interval $[0, \{w|\tau\} | \{W|\tau\} = \{w|\tau\}]$. And the conditional random variable $\{W|\tau\}$ plays a role identical to the unconditional W in Sec. 2. So, all the formulas for W in Sec. 2 can be applied to finding $\{W|\tau\}$. The only detail we must attend to, in order to find the PDF for total time as post-doc for those retained in the sample, is shifting the PDF for $\{W|\tau\}$ to the right of the origin by τ months, reflecting the conditioning information.

From Eq.(4), we can write,

$$f_X(x) = \frac{1 - F_{\{W|\tau\}}(x)}{E[\{W|\tau\}]} \quad x \geq 0$$

Then, from Eq.(5) and Eq.(6), we have

$$\hat{F}_{\{W|\tau\}}(x) = 1 - \frac{f_X(x)}{f_X(0)} \quad x \geq 0, \quad (5A)$$

and for the PDF,

$$f_{\{W|\tau\}}(x) = -\frac{1}{f_X(0)} \frac{d}{dx}(f_X(x)) \quad x \geq 0. \quad (6A)$$

Following Eqs. (7), (8) and (9), we can obtain the mean any one of three ways,

$$E[\{W|\tau\}] = \int_0^\infty [1 - F_{\{W|\tau\}}(x)] dx. \quad (7A)$$

$$E[\{W|\tau\}] = \int_0^\infty x f_{\{W|\tau\}}(x) dx. \quad (8A)$$

$$E[\{W|\tau\}] = \frac{1}{f_X(0)}. \quad (9A)$$

We have one final step to transform these results to depict the distribution and the mean of W_τ for all post-docs whose elapsed time as post-doc has at least been τ months. For the mean $E[W_\tau]$ total time as post-doc, including the τ months as post-doc needed to be in the sample, we must add back the excluded number of months,

$$E[W_\tau] = E[\{W|\tau\}] + \tau \quad (10)$$

For the PDF $f_{W_\tau}(w)$ of the random variable W_τ corresponding to the total time as post-doc, including the τ months as post-doc needed to be in the sample, we must use the PDF found for $\{W|\tau\}$ and shift it to the right by the deterministic number of months τ ,

$$f_{W_\tau}(x) = f_{\{W|\tau\}}(x - \tau), \quad \text{for } x \geq \tau \quad (11)$$

Eq. (11) gives us the PDF for the total time as post-doc for those included in the sample, including the τ months as post-doc that were required to be accepted into the sample. Since

the minimum allowable total time for being a post-doc is τ months, the PDF found with Eq. (11) will have value zero for $0 \leq x < \tau$, and start with its positive values at $x = \tau$.

4. Fixed Annual Dates for Both Post-doc Starts and the Survey

We now consider the case for which the starting time of becoming a post-doc *and* the time of survey sampling are *both* fixed during the course of a year, every year. For example, we may have all individuals in our sample becoming post-docs in the month of June, the traditional month of graduation, and the survey always occurs in the month of October, four months later. As usual, the survey asks the sampled individuals how long they have been post-docs. And, as usual, the true duration of a post-doc is a continuous non-negative random variable; it can terminate at any time. But with every-October sampling, we can only ‘see’ a post-doc in Octobers. If she terminates between two successive Octobers, she could not appear in the next sample. We use June for the date of both graduation and post-doc start as illustrative only, since post-doc positions can potentially start at any time. In reality, a large number are likely clumped in June and perhaps also September and January, whereas others are spread throughout the year. Later, we discuss how to model such situations.

Returning to our illustrative June–October example, let’s consider the situation from the point of view of the person conducting the survey. First, for that rare post-doc appointment that terminates quickly between June and October of the same year, he cannot be recorded in the sample; in a sense, he is invisible to the survey. The annual October survey will include responses from individuals who have been post-docs to-date for specific pre-specified amounts of time, in our example 4 months, $(4 + 12) = 16$ months, $(4 + 2*12) = 28$ months, 40 months, 52 months, etc. These times are deterministic, each successive time being one year later than the previous.

We can visualize the situation in the Temporal Sampling Diagram shown in Figure 4. The figure represents a stylized picture of the real system, assumed to be operating for a long time in steady state from the past and into the future with many more post-docs than can be shown in a space-constrained figure. As in Figure 1, time runs from left to right. The red vertical dashed lines correspond to the times each year that post-doc careers start, new careers starting at the same date each year. The dark blue vertical lines correspond to the times each year that current post-docs are sampled, always on the same date. The blue squares shown intercepting the post-doc careers depict those post-docs that are included in a given year’s sample. For example, four post-docs are intercepted in the second sample shown, taken at time t_2 , whereas seven post-docs are included in the following year’s sample at time t_3 .

We will be interested for each sampled post-doc in the number of previous sampling dates that have occurred to date in that post-doc’s career. For example, consider the sampling that occurs at time t_3 . Post-doc #3, the third from the top in Figure 4, when sampled at time t_3 , has experienced two previous sampling dates during her post-doc career. Post-doc #6 has experienced one previous sampling date, and post-doc #9 has experienced zero previous sampling dates. Note also from Figure 4 that there can be very short post-doc careers that

overlap no sampling dates, and thus are invisible to the sampling process; post-docs 5, 7 and 10 all fall into this category.

Note that we are only dealing with one-year sampling intervals and are unable to pinpoint post-doc durations to any more fine-grained detail. Our unit of time will be one year. As shown in Figure 4, we have a fixed delay time between moment of becoming a post-doc and time of the survey, $0 < \tau < 1$. In the discussion above with June post-doc starts and October surveys, that time was four months or $\tau = 1/3$ year. Note that τ must equal the fraction of the year between becoming a post-doc and the time of the *next* survey, a possible confusing issue if the survey occurs earlier in the year than the time of becoming a post-doc. For example, if individuals become post-docs in June and the annual survey occurs every February, we set $\tau = 2/3$ year, as there is an 8-month delay between post-doc starts and the next upcoming survey of post-docs.

To analyze the situation, we need to ask, “Is there a ‘random incidence’ sampling bias related to duration of post-doc career, a bias analogous to the one we found in Sec. 2?” And the answer is, “Yes!” Suppose for example that each June’s post-doc ‘class’ comprises $4*n$ individuals who are equally likely to have post-doc career durations between (a) 4 and 16 months, (b) 16 to 28 months, (c) 28 to 40 months, and (d) 40 to 52 months. And we assume that every June a new cohort of $4*n$ graduates, all becoming post-docs with the same equal probabilities over the four possible post-doc durations. Then, the sample survey in any given October will have n of type (a) post-docs, $2n$ for type (b) post-docs, $3n$ for type (c) post-docs, and $4n$ for type (d) post-docs. The n for type (a) post-docs is due to the fact that those whose post-doc duration is at least 4 months but less than 16 experience only one “October” as post-docs, and thus are eligible for the survey only once. For type (b) post-docs, the survey will include members not only of this year’s graduating class but last year’s as well. Type (b) post-docs are twice as likely as type (a) post-docs to arise in the sample. Each type-(b) post-doc experiences two Octobers while post-doc. And so on for types (c) and (d). The type (d) post-docs are four times as likely to arise in the sample as type (a), even though exactly n of each are produced every June. Thus, we see that the same ‘random incidence’ sampling bias due to post-doc durations is present in this situation having fixed times of starting post-docs and of sampling post-docs.

We now generalize this situation and first examine relative frequencies of post-docs whose eventual total durations fall into given time intervals. Define

$p_i \equiv$ fraction of newly created post-docs whose eventual post-doc duration falls into interval i , $i=0, 1, 2, 3, \dots$

The intervals begin with the fractional year of length τ then proceed in one-year increments, as follows:

Interval Designation	Duration of the Post-doc	Relative Frequency
$i = 0$	$[0, \tau)$	p_0

<u>Interval Designation</u>	<u>Duration of the Post-doc</u>	<u>Relative Frequency</u>
$i=1$	$[, 1+)$	p_1
$i=2$	$[1+ , 2+)$	p_2
.	.	.
.	.	.
.	.	.
$i= m$	$[(m-1)+ , m+)$	p_m
.	.	.
.	.	.
.	.	.

From our discussion above, we understand that post-docs whose eventual post-doc duration is in interval $i= m$ are, everything being equal, m times more likely to arise in the survey sample than those in interval $i=1$. Thus, as in Sec. 2, the likelihood of a post-doc occurring in the sample whose post-doc duration falls in the interval $i= m$ is directly proportional to two quantities: *the relative frequency p_m of such intervals and the index m of the interval*. The index m is a surrogate for the total duration of the post-doc career, representing in our case the total integer number of survey opportunities during the career. In fact, any post-doc whose eventual post-doc duration falls in the interval m has a true post-doc duration somewhere between $(m-1)+$ and $m+$ years.

Let us define our sampling relative frequencies,

$$\pi_m \equiv P\{\text{a type } m \text{ post- doc is drawn in the survey}\}, \quad m=0, 1, 2, \dots$$

From our previous arguments, we know that $\pi_m \propto mp_m$.

Let us now consider the integer random variable that depicts our situation,

$$\hat{S} \equiv \text{number of survey opportunities that arise during a random post- doc career.}$$

If, for instance, surveys occur during October, then the experimental value of $\hat{S}$ for any given post-doc would be the number of Octobers during that post-doc career. Experimental values for $\hat{S}$ are governed by the probabilities $p_0, p_1, p_2, \dots$. While $\hat{S}$ is related to W , note that $\hat{S} \neq W$. To compute W from $\hat{S}$, we need also to consider

$$\theta \equiv \text{fraction of year continuing as post- doc after last sampling, } 0 \leq \theta < 1.$$

As an example, if the last sampling event for the post-doc occurred in October of 2016 and the post-doc appointment terminated a half year later in April of 2017, then $\theta=1/2$. For any given post-doc who has been sampled and whose career has ended, the duration of her post-doc career W , can be expressed,

$$W = (\hat{S} - 1) + \Delta + \theta. \quad (12)$$

Example—If only one sampling event occurred during the post-doc career (i.e., $\hat{S}=1$), then the total duration of the career is equal to the sum of Δ , the post-doc time until that sampling, plus θ , the fraction of the year left as post-doc after that sampling time. The term $(\hat{S}-1)$ adds the appropriate integer number of years for longer duration post-doc careers.

For sampled post-docs we see that the random variable $\hat{S}$ can assume any positive integer value, 1, 2, ... (The fractional year Δ is assigned a value of zero since there are zero sampling opportunities in a short post-doc career of duration less than Δ .) We can express the mean of $\hat{S}$ as follows,

$$E[\hat{S}] = \sum_{m=0}^{\infty} mp_m. \quad (13)$$

We can now write for the sampling relative frequencies,

$$\pi_m = \frac{mp_m}{E[\hat{S}]}, \quad m=0, 1, 2, \dots \quad (14)$$

Eq. (14) is analogous to Eq.(2) in Sec. 2.

We now develop an analogue for Eq.(3). Suppose we sample an individual post-doc whose eventual experimental value for $\hat{S}$ is s_0 , where $s_0 = 1, 2, 3, \dots$. We are equally likely to perform our sampling at any of the s_0 surveys over her entire post-doc career duration. Hence we know that the cumulative number of sampling opportunities to date while a post-doc is an integer random variable uniformly distributed over the interval $[1, s_0]$. More formally, if $P_{X|\hat{S}}(x | s_0)$ is the conditional probability that $X = x$, given that $\hat{S} = s_0$, then

$$P_{X|\hat{S}}(x | s_0) = \frac{1}{s_0}, \quad x=1, 2, \dots, s_0. \quad (15)$$

Following familiar logic, we find an analogue to Eq.(4). We write an expression for $P_X(x)$, the unconditional probability that for a randomly selected post-doc the cumulative number of sampling opportunities to date while a post-doc assumes an experimental value x ,

$$P_X(x) = \sum_{s_0=x}^{\infty} \frac{1}{s_0} \frac{s_0 p_{s_0}}{E[\hat{S}]} = \sum_{s_0=x}^{\infty} \frac{p_{s_0}}{E[\hat{S}]} = \frac{1 - F_{\hat{S}}(x-1)}{E[\hat{S}]}, \quad x=1, 2, 3, \dots, \quad (16)$$

where $F_{\hat{S}}(x)$ is the CDF of the random variable $\hat{S}$, $F_{\hat{S}}(x) \equiv \Pr\{\hat{S} \leq x\}$, for $x=1, 2, 3, \dots$

In examining Eq.(16), we see that $P_X(x)$ is directly proportional to the CCDF having shifted argument $(x-1)$, i.e., $[1 - F_{\hat{S}}(x-1)] \equiv \Pr\{\hat{S} > x-1\}$, $x=1, 2, 3, \dots$

This means that $P_X(x)$ is a monotonically non-increasing function of its argument, for $x=1, 2, 3, \dots$. In simple terms, this probability mass function (PMF) starts at its maximum value at $x=1$, and eventually declines from there to 0 as x increases.

Since $P_X(x)$ is directly proportional to $[1 - F_{\hat{S}}(x-1)] \equiv \Pr\{\hat{S} > x-1\}$, we simply normalize the large sample histogram found for the elapsed integer time X in those sampled from the post-doc population. The normalization will make the area under the step-valued histogram equal to 1.0, a requirement for any PMF. Then observe that the unit-shifted CCDF for the desired random variable $\hat{S}$, $[1 - F_{\hat{S}}(x-1)]$, is directly proportional to normalized histogram. This gives us the shape of the desired CCDF. Since $F_{\hat{S}}(1-1) = F_{\hat{S}}(0) = 0$, we must have $[1 - F_{\hat{S}}(0)] = 1$. This suggests that the proportionality relationship can be rewritten as an equation,

$$[1 - F_{\hat{S}}(x-1)] = \frac{P_X(x)}{P_X(1)}, \text{ assuming that } P_X(1) > 0, \text{ which it must be by the monotonicity property of } P_X(x), \quad x=1, 2, 3, \dots$$

We now arrive at analogues for Eqs.(5) and (6),

$$F_{\hat{S}}(x-1) = 1 - \frac{P_X(x)}{P_X(1)} \quad x=1, 2, \dots$$

Setting $y \equiv x-1$, we have the desired relationship,

$$F_{\hat{S}}(y) = 1 - \frac{P_X(y+1)}{P_X(1)} \quad y=0, 1, 2, \dots, \quad (17)$$

and for the PMF,

$$\begin{aligned} P_{\hat{S}}(s_0) &\equiv \Pr\{\hat{S}=s_0\} = F_{\hat{S}}(s_0) - F_{\hat{S}}(s_0-1) \\ &= \left[1 - \frac{P_X(s_0+1)}{P_X(1)}\right] - \left[1 - \frac{P_X(s_0)}{P_X(1)}\right], \quad s_0=1, 2, 3, \dots \end{aligned}$$

which simplifies to

$$P_{\hat{s}}(s_0) = \frac{P_X(s_0) - P_X(s_0+1)}{P_X(1)}, \quad s_0 = 1, 2, 3, \dots \quad (18)$$

We know that the numerator must be non-negative due to the downward monotonicity of $P_X(s_0)$.

These are the transformations we are looking for: Deriving the probability law for $\hat{S}$, the number of sampling opportunities as a post-doc only from the histogram of integer cumulative number of sampling opportunities to date of sampled post-docs.

Again here there is a simple way to compute the mean $E[\hat{S}]$. Since $P_X(x) = \frac{1 - F_{\hat{s}}(x-1)}{E[\hat{S}]}$, $x = 1, 2, 3, \dots$ and $F_{\hat{s}}(1-1) = 0$, we have $P_X(1) = \frac{1}{E[\hat{S}]}$, or

$$E[\hat{S}] = \frac{1}{P_X(1)} \quad (19)$$

We can now invoke Eq.(12) to obtain an estimate of the true mean post-doc career length, W ,

$$E[W] = E[(\hat{S}-1) + \Delta + \theta] = E[(\hat{S}-1)] + \Delta + E[\theta]$$

If we assume the post-doc ends her post-doc career approximately $\frac{1}{2}$ year following the last survey event⁴, then we have,

$$E[W] = \left(\frac{1}{P_X(1)} - 1\right) + \Delta + 1/2,$$

or

$$E[W] = (\Delta - 1/2) + \frac{1}{P_X(1)} \quad (20)$$

⁴We have more accurate ways to estimate $E[\theta]$. At this time, we believe that in most circumstances, $1/3 < E[\theta] < 1/2$.

5. Illustrative Example, with Inclusion of Integer Round Off Issues

Here we present a numerical example for the sampling assumptions of Section 2, “*Case 1: Random Career Start Times, Fixed Survey Time.*”

5.1. Discretized Data

We may expect to obtain responses in discrete “data buckets,” where we obtain the number who have reported to have been post-docs 0 to 1 year, 1 to 2 years, 2 to 3 years, etc., being careful to handle in a statistically appropriate way any respondent answers that are truncated to the nearest integer. The discrete time unit may be less than a year, say a month or even a week. Whatever the discrete time unit of measurement, we must be aware of integer round off tendencies and ways to handle them. We offer two solutions: A quick and easy one, and a slightly more complex one.

A Simple Procedure—For situations in which all respondents provide only integer answers, in years, we can “probabilistically distribute” those answers within adjacent unit-length intervals. For example, for those who answer 2 years, one can argue that about 50% of those responses should be placed randomly in the right half of the 1-to-2-year bucket and the other 50% placed randomly in the left half of the 2-to-3-year bucket. The logic behind this: There are likely to be roughly as many post-docs who have spent 1.8 years and who round to 2 as there are post-docs who have spent 2.2 years and who also round to 2. Under this scheme, any answer of zero years would be placed randomly in the interval $(0, \frac{1}{2})$.

A More Complex Procedure—Now consider the aggregate table of results shown in Table 2. Column 2 of the table (labeled “Original”) is a stylized version of a large sample with 8,685 post-doc respondents over a multi-year period, reported by the National Science Foundation⁵. We will work on the stylized data to develop a heuristic procedure for dealing with integer round off issues. (Note: The NSF survey provides data rounded to the nearest month and includes only post-docs who are at least 18 months from completion of the PhD. Our manipulation of “stylized data,” rounded to years, is for illustrative purposes only, and should not be confused a detailed statistical analysis of the NSF survey. That is the subject of another paper.)

The n^{th} row of the table gives the numbers of post-docs who reported to have already been active post-docs for $n-1$ to n years, $n=1, 2, 3, \dots$. With the exception of row 1, we see that the aggregate figures drop monotonically as n increases from $n=2$ to $n=3$, to $n=4$, etc.

Our theoretical results earlier in this paper require that *all* entries in this table must be monotonically decreasing, starting with “0 to 1 year” being the maximum value. One can see clearly from the Table 2 data that this is not the case. For large sample sizes such as these, this cannot be, if everyone is reporting properly.

We propose one possible explanation for this data anomaly. It is that, despite instructions, many people reporting results about themselves round to the nearest integer. Someone who

⁵<http://www.nsf.gov/statistics/srvydoctoratework/>

has been a post-doc for 10 months may round up to ‘1 year’, just as someone who has been a post-doc for 14 months may round down to 1 year. This rounding may be due to intellectual laziness, or perhaps due to the respondent’s not remembering accurately the answer to the question: “*How long have you been a post-doc?*” Or maybe the start date of the post-doc contract was vague in some sense.

Here is our hypothesis—The integer-based data buckets tend to create a bias, if respondents report in this way -- often rounding to the nearest integer. Suppose the data bucket for “Less than one year” is a unit-length interval, closed on the left and open on the right: $[0, 1)$. That is, only entries strictly less than 1.0 are included. Then the person who has been a post-doc for 10 months and rounds to 1 year is incorrectly placed in bucket #2, closed on the left and open on the right: $[1, 2)$. For higher-level buckets this error approximately washes out, as those buckets gain some respondents from below and lose respondents to above. For example, the interval $[1, 2)$ gains integer round-ups from $[0, 1)$ and loses due to its own integer round-ups to $[2, 3)$. Since $[1, 2)$ is closed on the left, any integer round-downs from within that interval have no effect on the count of respondents in the interval, as they are still seen as within the interval. Note that $[0, 1)$, the left-most interval, is the only interval having zero round-ups from its left. Interval $[0, 1)$ only loses respondents; it has no way to gain them.

If this data bucket coding procedure is found to occur, we would need to find a “correction algorithm” that approximately corrects for the inherent bias. We propose one such method below.

A Proposed Correction Algorithm—We present our correction algorithm with a fixed given assumption about individual’s propensity to round off to the nearest integer. After going through the procedure one can see readily how to adapt the method to any empirically derived model for how people round off to nearest integers.

We conjecture that in any integer-length interval, the probability of reporting the closest integer instead of the true value rises linearly from zero to 1.0, starting at the midpoint of the interval and moving each way, down to the next lower integer and up to the next higher integer. This is shown in Figure 5, modeled for post-docs who have been active post-docs for a period ranging from 1 to 2 years. If the model that is depicted in Figure 5 represents reality, and if we assume that the true distribution of times spent as post-docs in this interval is *uniformly distributed over the entire interval*, then fully 25% of the respondents round down to 1 year and another 25% of respondents round up to 2 years.

To see how this creates bias, consider the interval from 0 to 1 year. We doubt that any respondents having less than six months as a post-doc would round down to 0 years, so the left-hand probability straight line is most likely absent in this case, the only interval with this property. And even if they do round down to zero, it does not affect our results. But the right hand probability line would still be active, meaning that fully 25% of those within the 0 to 1 year window would report spending 1 year as a post-doc. If the data buckets are closed intervals on the left and open on the right, meaning that the first data bucket is the interval $[0, 1)$, then all of those responses that are rounded up to 1 would be counted in the next

higher $[1, 2)$ bucket. If this is the case, then the number of respondents shown in the table for $[0, 1)$ is only 75% of the true number of respondents. To correct the situation, we need to add another one-third of the current number shown for the $[0, 1)$ interval. We can obtain this additional number of respondents by taking them from the number shown in the next higher interval, $[1, 2)$. So, in Step 1 of the algorithm we take one-third of the current number shown for $[0, 1)$ from the next higher interval, $[1, 2)$. In Step 2 of the algorithm we do the same thing for intervals $[1, 2)$ and $[2, 3)$. That is, after we obtain the new reduced entry for interval $[1, 2)$ we argue that we need to add another one-third of that number, taking it from the adjacent $[2, 3)$ interval. That is, we take one-third of the new number shown in $[1, 2)$ from $[2, 3)$. This process continues until we are finished. The algorithm will have n steps, one for each unit-length interval in the data set. Note that the ‘big beneficiary’ of this process is the first interval, $[0, 1)$. The others lose to the adjacent interval to the left but gain somewhat from the adjacent one to the right. Due to downward monotonicity, the gain from the right is likely to be less than the loss to the left in each instance.

Column 3 of Table 2, labeled “Final”, contains the results of applying the algorithm to the stylized NSF data. Note that monotonicity is now restored, starting with the first interval, $[0, 1)$.

One could make this method more precise by using an empirically based “rounding up probability curve” to improve and replace the linear probability curve we postulated here for intuitive appeal and ease of use. Then, continuing to use the assumption of a conditionally uniform probability density function over each unit-length interval, one could determine for each interval the fraction of respondents from within the interval that round up and thus move into the next higher interval.

5.2. The Illustrative Numerical Example

We now proceed to our illustrative numerical example, using the modified stylized data entries of column 3 of Table 2.

Step 1. Our illustrative sample of post-docs contains $N=8,685$ entries. Since this is a large sample size with monotonicity in place with the data, we see no need to ‘smooth’ the data. However, we need to deal with the aggregated right-hand tail of the distribution, where 76 sampled post-docs report having spent more than (an incredible) 12 years as post-docs. With over 8,000 in the sample, virtually any reasonable way of dealing with the right-hand tail will suffice, as the net consequences will be in the statistical noise. One way to do this is to deploy 13 post-docs to the interval 12-to-13, another 13 to 13-to-14, another 13 to 14-to-15, another 13 to 15-to-16, another 13 to 16-to-17, and the final 11 to 17-to-18. That deployment yields the least total time spent as post-docs, since we need to preserve unimodality in the estimated PDF. That is, we cannot allow any subsequent entry to be larger than the previous one. The resulting histogram is displayed in Figure 6. Note that what we have here is an estimated **probability mass function** (PMF), not a PDF – the latter associated with strictly continuous random variables and the former with discrete (usually integer) random variables. Our need for discrete random variables in this instance is due to the way the data were collected, in one-year “data buckets.”

Step 2. The PMF can also be considered to be a piece-wise flat PDF, with constant value between successive integers. The mean of this empirically derived PMF (or PDF) is the estimated average time spent to date as a post-doc. Let's call it $\hat{E}[X]$. In doing the calculations, we assign the mid point of each integer interval to the sum for computing the mean. Plus we assign a conditional mean of 14.5 to the 76 data points comprising the far right-hand tail. We compute $\hat{E}[X] = 0.5(0.315) + 1.5(0.306) + 2.5(0.165) + \dots = 2.147$ years.

Step 3. Here we are to identify the numerical value for $\hat{f}_X(0)$, which is the maximum value for the function $\hat{f}_X(x)$ for all $x \geq 0$. Following the PMF-to-PDF logic above, we do this by equating the height of the PDF to be equal to the value of the PMF at each respective integer value. The estimated PDF is a sequence of rectangles, the height of each one equal to the numerical value of the corresponding PMF value. As an example, $\hat{f}_X(x)$ for $2 < x \leq 3$ is equal to 0.1656 (from Table 2 and as shown in Figure 6). In a like way, we find the desired quantity for Step 3: $\hat{f}_X(0) = 0.3164$.

Step 4. Now we find the desired estimated CDF $\hat{F}_W(x)$ of the random variable W ,

the total duration of the post-doc status, $\hat{F}_W(x)$, using Eq.(5), $\hat{F}_W(x) = 1 - \frac{\hat{f}_X(x)}{\hat{f}_X(0)}$.

Substituting $\hat{f}_X(0) = 0.3164$, we have $\hat{F}_W(x) = 1 - \frac{\hat{f}_X(x)}{0.3164}$. We see that $\hat{F}_W(x) = 0$ for

$0 \leq x < 1$, $\hat{F}_W(x) = 1 - \frac{0.3077}{0.3164} = 0.028$ for $1 \leq x < 2$, $\hat{F}_W(x) = 1 - \frac{0.1656}{0.3164} = 0.476$ for $2 \leq x < 3$, etc. The entire CDF is shown in Figure 7. The CCDF is shown in Figure 8.

From Figure 5, we can obtain the estimated mean of W . Let's call it $\hat{E}[W]$. Invoking Eq.(7), we can write

$$\hat{E}[W] = \int_0^\infty [1 - F_W(x)] dx = \sum_{n=1}^{\infty} [1 - F_W(n)] = 1 + 0.972 + 0.524 + \dots = 3.175 \text{ years}$$

Note that $3.175 = \hat{E}[W] > \hat{E}[X] = 2.174$ years, a result quite different from a system in which post-doc careers are governed by a negative exponential PDF.

Step 5. Due to the integer discreteness of the random variables, we need not 30 differentiate to find a PDF, which we already know.

Step 6. The issues involving integer bucket data were addressed above.

6. Summary and Discussion

We have developed methods to estimate accurately the distribution of total time spent as post-docs knowing only a histogram of time spent to date as a post-doc for a random snapshot sample of post-docs. Beyond post-docs, the methods apply to any application in which one is sampling individuals who are in some temporary status and where one wishes to project distributionally the total time that they will be in that status. We derived the results for various cases: (1) where either the start times of post-docs or the times of the survey sampling were random (or both were random); (2) under condition (1), where only post-docs who had served some minimal pre-specified time as post-docs were included in the sample; and (3) where both the time of starting post-docs and the time of the survey were

deterministically fixed at given times in the year. Again, these results are general, applicable beyond post-docs.

We expect that in data-intensive applications some of the assumptions of this paper will have to be modified to work properly in practice. In many cases this should be straightforward. For instance, suppose there is a heterogeneous population of post-docs, 25% of whom start post-docs in January, 25% in June, 25% in September and the remaining 25% whose start times can be viewed as totally random over the course of a year. And also suppose that the survey is always done in October. One can divide this problem into four parts, applying the theory of Sec. 4 of the paper three separate times for the first three sub-cohorts and applying Sec. 2 of the paper to the fourth sub-cohort. Then, one simply merges the results so obtained. We would expect somewhat “bumpy” distributional results as Sec. 4 only offers results by the year whereas Sec. 2 can be much more fine-grained, say to the month or week.

Finally, suppose the sub-population being surveyed can be partitioned into K groups, where K is a positive integer. If for example, $K = 2$, we might partition by gender, male or female. If $K = 5$, we might partition by five separate socio-economic classes. For each partitioned segment we can use the results of this paper to find the probability law for the total duration for that segment in its temporary status. Then, by doing this for each partitioning segment, one can compare durations across segments. For example, with a dataset of post-docs, one can examine whether the post-doc career durations depend on gender, country of birth, socio-economic classification or any one of a number of different attributes.

For readers interested in following more of the post-doc analysis, we cite the aforementioned companion paper by the same authors, “The Post-doc Queue: A Labor Force in Waiting” (under review, 2016). Finally, incorporating all three sampling settings of this paper, the same authors are preparing a detailed statistical analysis of the actual NSF doctoral survey (no stylized data): “Using NSF Doctoral Surveys to Infer Total Durations of Post-doc Careers (in preparation).

Acknowledgments

The National Institute of General Medical Sciences and the Office of Behavioral and Social Sciences Research of the National Institutes of Health (NIH) supported this work (Grant 2U01GM094141-05). The grant, ‘A Model-Based Examination of Behavioral & Social Science Workforce: Improving Health Outcomes,’ was awarded to the Ohio State University, MIT and Virginia Tech. The discussion and conclusions in this paper are those of the author and do not necessarily represent the views of the National Institutes of Health, the Ohio State University, MIT or Virginia Tech. For detailed comments on an earlier draft we sincerely thank Ed Kaplan and Navid Ghaffarzadegan.

References

- Academy Health, Health Services Research Methods. <http://www.hsrmeths.org/Glossary/Terms/L/Length%20Biased%20Sampling.aspx>. Retrieved Sept. 4, 2015
- Andalib, Maryam, Ghaffarzadegan, Navid, Larson, Richard. The Post-doc Queue: A Labor Force in Waiting. 2016 under review.
- Andalib, Maryam, Ghaffarzadegan, Navid, Larson, Richard. Using NSF Doctoral Surveys to Infer Total Durations of Post-doc Careers. 2016 in preparation.
- Barnett, Karen, Mercer, Stewart W., bury, Michael Nor, Watt, Graham, Wyke, Sally, Guthrie, Bruce. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*. 380:37–43. Published online 10 May 2012.

- Blumstein, Alfred, Cohen, Jacqueline. Characterizing Criminal Careers. *Science*. Aug 28.1987 :985–991.
- Cox, DR. *Renewal Theory*. Methuen; London: 1962.
- Drake, AW. *Fundamentals of Applied Probability*. 2. McGraw-Hill; New York: 1967.
- Laidra, Kaia, Pullmann, Helle, Allik, Jüri. Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences*. 2006; doi: 10.1016/j.paid.2006.08.001
- Larson, Richard C., Odoni, Amedeo R. *Urban Operations Research*. Prentice Hall; Englewood Cliffs, NJ: 1981.
- Kaplan, Edward H. Snapshot Samples. *Socio-Economic Planning Sciences*. 1997; 31(4):281–291.
- Oum, Tae Hoon. A Cross Sectional Study of Freight Transport Demand and Rail-Truck Competition in Canada. *The Bell Journal of Economics*. Autumn;1979 10(2):463–482.
- Phelan, Jo C., Link, Bruce G. The Illusion of Simple Truths: How Current Status Bias Distorts Perceptions of Homelessness and Other Social Problems. Unpublished draft. 1996. Retrieved from Internet <http://fla.st/1JNpKmw> Sept. 2, 2015
- Rindfleisch, Aric, Malter, Alan J., Ganesan, Shankar, Moorman, Christine. Cross-Sectional Versus Longitudinal Survey Research: Concepts, Findings, and Guidelines. *Journal of Marketing Research*. 2008 Jun; 45(3):261–279.
- Tunçalp Ö1, Souza JP, Hindin MJ, Santos CA, Oliveira TH, Vogel JP, Togoobaatar G, Ha DQ, Say L, Gülmezoglu AM. WHO Multicountry Survey on Maternal and Newborn Health Research Network. *PubMed BJOG*. 2014 Mar; 121(Suppl 1):57–65. [accessed Sept. 2, 2015] <http://www.ncbi.nlm.nih.gov/pubmed/24641536>. DOI: 10.1111/1471-0528.12634
- Voleti, Sudhir, Bharadwaj, Sundar. [accessed Sept. 2, 2015] Bayesian Estimation and Clustering of Latent Attitudinal Parameters Using Cross-sectional Survey Data: Application to an Online Banking Survey. *Indian School of Business Research Paper Series*. Date posted: February 10, 2015 <http://ssrn.com/abstract=2562187>

Biography

Richard C. Larson received his BS and PhD degrees in Electrical Engineering from MIT where is Mitsui Professor in MIT's Institute for Data, Systems and Society. Prof. Larson is author of over 100 scientific articles, primarily in the fields of operations research as applied to service industries, health care, technology-enabled education, urban service systems, queueing, logistics and workforce planning. He has served as President of ORSA, (1993–4), is Past-President of INFORMS, Institute for Operations Research and the Management Sciences, and has served as consultant to the World Bank, the United Nations, the City of New York and numerous private companies. Prof. Larson's research on queues has not only resulted in new computational techniques, but has also been covered extensively in national media. Prof. Larson has served as Co-Director of the MIT Operations Research Center (over 15 years in that post). He is a member of the National Academy of Engineering and is an

INFORMS Founding Fellow. He has been honored with the INFORMS President's Award, Lanchester Prize (for best publication in one year) and the Kimball Medal.

Prof. Larson is Founding Director of LINC, Learning International Networks Consortium, an MIT-based international technology-enabled education project that will hold its next international symposium in spring 2016. He is Principal Investigator of MIT BLOSSOMS, *Blended Learning Open Source Science or Math Studies* <http://blossoms.mit.edu>, focused on high school STEM education (STEM = Science, Technology, Engineering and Mathematics) and STEM teacher Professional Development. This is an international collaboration with ten countries. Many of these high school lessons focus on health care and disease prevention, including diabetes II, pandemic influenza, cancer and diseases of the eye.

Dr. Larson's current MIT research includes modeling of infectious diseases, especially pandemic influenza and MERS; post-traumatic stress disorder; K-12 STEM education as a complex system; home energy management; and Ph.D.-level workforce planning for the NIH.

Highlights

We derive a statistical procedure for researchers to infer the probability law of the total time that a person will be in some temporary state by asking a representative “one-shot” survey sample of such persons how long they have already been in that state.

We develop results for several types of cross-sectional or one-shot sampling, including

- Random and fixed times of surveying;
- Random and fixed times of entering the state; and
- Sampling only those who have already spent some minimal specified time in the targeted state.

An example with post-doc data is included to demonstrate the steps.

Figure 1.
Sample-based intercepting of post-doc careers at a random time.

Figure 2.
Temporal Sampling Diagram of Case 1, Random Career Start Times, Fixed Survey Time:
Relationship Between W – the Eventual Total Length of Time as a Post-doc and X - the Elapsed Time Spent as a Post-doc at Time of Survey

Figure 3.

Temporal Sampling Diagram for the τ -Delayed Case. Relationship Between W_τ – the Eventual Total Length of Time as a Post-doc, given that this time must exceed some pre-specified minimum time τ , $\{W|\tau\}$ -- the conditional total time as a post-doc after time τ , and X - the Elapsed post- τ Time Spent as a Post-doc at Time of Survey.

Figure 4. Temporal Sampling Diagram. The case of sampling post-docs when yearly start dates of post-docs are fixed and yearly sampling dates are fixed. ● corresponds to start dates of post-doc careers. ■ corresponds to post-doc careers intercepted by a given survey.

Figure 5.
A Respondent's Probability of Rounding to Nearest Integer

Figure 6. Histogram of $\hat{f}_X(x)$, the Data-Estimated PDF for X , the Elapsed Time in as a Post-doc to Date

Figure 7.

$F_W(x)$: The CDF of W – the total length of time as a post-doc – for the numerical example.

Figure 8.
[$1 - F_W(x)$]: The CCDF (Complementary Cumulative Distribution Function) for W , for the numerical example.

Table 1

Summary of Technical Terminology

W	Random variable denoting the total duration of time in temporary state (e.g., duration of a post-doc career for a randomly starting post-doc).
$f_W(w)$	Probability density function (PDF) for random variable W .
$F_W(x)$	Cumulative distribution function (CDF) for random variable W .
$E[W]$	Expected value of random variable W .
X	Random variable denoting the elapsed time to date in temporary state (e.g., elapsed career time to date of post-doc of a randomly sampled post-doc).
$f_X(x)$	PDF for random variable X .
τ	The minimal amount of time a sampled person must have already been in temporary state (e.g., post-doc) to be included in the final sample.
	Fixed delay time between moment of entering temporary state (e.g., becoming a post-doc) and time of the cross-sectional survey.
θ	Fraction of a year after last sampling until conclusion of time in temporary state (e.g., being a post-doc)

Table 2

Histogram Data for Unit-interval Time Already Spent as a Post-doc

Duration	Original	Final	Fraction
$0 < t \leq 1$	2,051	2735	0.3164
$1 < t \leq 2$	2,679	2660	0.3077
$2 < t \leq 3$	1,739	1432	0.1656
$3 < t \leq 4$	917	745	0.0862
$4 < t \leq 5$	522	448	0.0518
$5 < t \leq 6$	318	275	0.0318
$6 < t \leq 7$	165	128	0.0148
$7 < t \leq 8$	86	72	0.0083
$8 < t \leq 9$	49	41	0.0047
$9 < t \leq 10$	36	34	0.0039
$10 < t \leq 11$	28	26	0.0030
$11 < t \leq 12$	16	13	0.0015
> 12	79	76	0.0088
	8,685	8,685	